ESTATÍSTICA APLICADA

Teste Qui-Quadrado

O Qui Quadrado, simbolizado pela letra grega *chi* elevada ao quadrado (χ^2), é como o teste t de Student: um teste de hipótese. Você se lembra que para o teste t usávamos a ***média*** e o *desvio-**padrão*** (**parâmetros**) de duas amostras para compararmos? No caso do teste qui-quadrado não usamos esses parâmetros, sendo portanto um **teste não-paramétrico**.

É um teste não-paramétrico que possibilita a gente ver a associação entre variáveis **qualitativas**.

O princípio básico do teste é comparar as **proporções**. Para isso são comparadas as divergências entre as frequências ***observadas*** (que foram obtidas da nossa amostra) e aquelas ***esperadas*** para esse evento.

Esperadas? Vamos dar um exemplo:

Em 1908 o matemático inglês Godfrey H. Hardy e o médico alemão Wilhem Weinberg concluíram que, se nenhum fator evolutivo atuasse sobre uma população que satisfizesse certas condições, as frequências de seus alelos permaneceriam inalteradas ao longo das gerações. Esse princípio ficou conhecido como lei ou teorema de Hardy-Weinberg ou princípio do equilíbrio gênico.

Suponhamos uma população em equilíbrio gênico, na qual há dois alelos, **A** e **a**, não ligados ao sexo. Sabendo-se que cada gameta porta apenas um alelo de cada gene, conclui-se que há apenas três possibilidades para a progênie de indivíduos heterozigotos para esse alelo: **AA**, **Aa** e **aa**, cujas probabilidades são respectivamente de ¼, ½ e ¼.

Em uma amostra, 100 descendentes foram estudados, encontramos as informações a seguir:

Genótipo	AA	Aa	aa
Frequência Observada	26	45	29

Esses seriam os valores ***observados***.

Quais seriam os valores ***esperados***? Mantendo a proporção de 25%, 50% e 25%, em uma amostra de 100 pessoas, teríamos:

Genótipo	AA	Aa	aa
Frequência Esperada	25	50	25

- **Como calcular**

Karl Pearson propôs a seguinte fórmula para medir as possíveis discrepâncias entre proporções ***observadas*** e ***esperadas:***

$$\chi^2 = \sum[(o-e)^2/e]$$

onde:

o: frequência observada para cada classe.

e: frequência esperada para cada classe.

Perceba que as frequências ***observadas*** são obtidas diretamente dos dados das amostras, enquanto que as frequências ***esperadas*** são calculadas a partir destas.

É importante notar que o desvio d = (o - e) é a diferença entre a frequência observada e a esperada em uma classe. Quando as frequências observadas são muito próximas às esperadas, o valor de χ^2 é pequeno. Mas, quando as divergências são grandes (o - e) passa a ser também grande e, consequentemente, χ^2 assume valores altos.

- **Hipóteses a serem testadas**

Como é um teste de hipótese, o teste qui-quadrado (χ^2) também trabalha com duas hipóteses:

Hipótese nula: As frequências observadas **não são diferentes** das frequências esperadas. Não existe diferença entre as frequências (contagens) dos grupos.

Hipótese alternativa: As frequências observadas são diferentes das frequências esperadas, portanto existe diferença entre as frequências.

- **Procedimento**

Calcule as diferenças, divida-as pelo número esperado, depois some os valores, resultando no Qui-Quadrado total

Genótipo	AA	Aa	aa	TOTAL
Frequência Observada	26	45	29	100
Frequência Esperada	25	50	25	100
Diferenças	26-25 = 1	45-50 = -5	29-25 = 4	0
Diferenças²	1^2 =1	$(-5)^2$ =25	4^2 =16	42
Diferenças²/Frequência Esperada	1/25 0.04	25/50 0.5	16/25 0.64	χ^2= 1.18

Logo, o χ^2calculado=$\sum(d^2/e)$= 1.18

É necessário obter duas estatísticas denominadas **χ^2calculado** e **χ^2 tabelado**. O χ^2 calculado é obtido pela fórmula e o χ^2 tabelado é obtido nas tabelas existentes. No final desta apostila existe um exemplo.

Se χ^2calculado > ou = χ^2 tabelado: *Rejeita-se* Ho.
Se χ^2 calculado < χ^2 tabelado: *Aceita-se* Ho.

Para obtermos o χ^2 tabelado usamos dois fatores essenciais: os **graus de liberdade** (g.l. ou em inglês *degrees of freedom* – d.f.) e o valor de **alfa** (α), que é o nosso nível de significância (igual ao teste t). Tradicionalmente, adota-se o nível de significância de 5%. Para calcularmos os graus de liberdade, usamos para esse ***TESTE DE ADERÊNCIA***:

G.L. = número de classes - 1

Nesse caso, temos 3 classes → g.l. = 2

- ***Como interpretar o resultado?***

No exemplo dado, como o valor de Qui Quadrado obtido (*1.18*) para as 03 classes (g.l. = 2) foi ***menor*** que o esperado ao acaso (α=0.05, χ^2tabelado = 5.99), ***não rejeita a hipótese nula*** e admitimos que as frequências esperadas dos genótipos são iguais às frequências observadas, portanto, essa amostra segue o equilíbrio de Hardy-Weinberg.

Tabela de Distribuição de χ^2.

Chi Square Distribution Table

d.f.	$\chi^2_{.25}$	$\chi^2_{.10}$	$\chi^2_{.05}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$	$\chi^2_{.001}$
1	1.32	2.71	3.84	5.02	6.63	7.88	10.8
2	2.77	4.61	5.99	7.38	9.21	10.6	13.8
3	4.11	6.25	7.81	9.35	11.3	12.8	16.3
4	5.39	7.78	9.49	11.1	13.3	14.9	18.5
5	6.63	9.24	11.1	12.8	15.1	16.7	20.5
6	7.84	10.6	12.6	14.4	16.8	18.5	22.5
7	9.04	12.0	14.1	16.0	18.5	20.3	24.3
8	10.2	13.4	15.5	17.5	20.1	22.0	26.1
9	11.4	14.7	16.9	19.0	21.7	23.6	27.9
10	12.5	16.0	18.3	20.5	23.2	25.2	29.6
11	13.7	17.3	19.7	21.9	24.7	26.8	31.3
12	14.8	18.5	21.0	23.3	26.2	28.3	32.9
13	16.0	19.8	22.4	24.7	27.7	29.8	34.5
14	17.1	21.1	23.7	26.1	29.1	31.3	36.1
15	18.2	22.3	25.0	27.5	30.6	32.8	37.7
16	19.4	23.5	26.3	28.8	32.0	34.3	39.3
17	20.5	24.8	27.6	30.2	33.4	35.7	40.8
18	21.6	26.0	28.9	31.5	34.8	37.2	42.3
19	22.7	27.2	30.1	32.9	36.2	38.6	32.8
20	23.8	28.4	31.4	34.2	37.6	40.0	45.3
21	24.9	29.6	32.7	35.5	38.9	41.4	46.8
22	26.0	30.8	33.9	36.8	40.3	42.8	48.3
23	27.1	32.0	35.2	38.1	41.6	44.2	49.7
24	28.2	33.2	36.4	39.4	32.0	45.6	51.2
25	29.3	34.4	37.7	40.6	44.3	46.9	52.6
26	30.4	35.6	38.9	41.9	45.6	48.3	54.1
27	31.5	36.7	40.1	43.2	47.0	49.6	55.5
28	32.6	37.9	41.3	44.5	48.3	51.0	56.9
29	33.7	39.1	42.6	45.7	49.6	52.3	58.3
30	34.8	40.3	43.8	47.0	50.9	53.7	59.7
40	45.6	51.8	55.8	59.3	63.7	66.8	73.4
50	56.3	63.2	67.5	71.4	76.2	79.5	86.7
60	67.0	74.4	79.1	83.3	88.4	92.0	99.6
70	77.6	85.5	90.5	95.0	100	104	112
80	88.1	96.6	102	107	112	116	125
90	98.6	108	113	118	124	128	137
100	109	118	124	130	136	140	149

Table from Ronald J. Wonnacott and Thomas H. Wonnacott, *Statistics: Discovering Its Power*, New York: John Wiley and Sons, 1982, p.352.

Exercício 1

Segundo Mendel, os resultados dos cruzamentos de ervilhas amarelas redondas com ervilhas verdes enrugadas seguem uma distribuição de probabilidades dada pela seguinte razão de fenótipos: 9:3:3:1.

Uma amostra de 556 ervilhas resultantes de cruzamentos de ervilhas amarelas redondas com ervilhas verdes enrugadas foi classificada (tabela a seguir). Há evidências de que os resultados desse experimento estejam de acordo com a distribuição de probabilidades proposta por Mendel? Considere um $\alpha = 0{,}05$e a tabela de Chi Quadrado fornecida acima.

Resultado	Amarela Redonda	Amarela Enrugada	Verde Redonda	Verde Enrugada
Frequência Observada	315	101	108	32

Exercício 2

Deseja-se verificar se o número de acidentes em uma estrada muda conforme o dia da semana. O número de acidentes observado para cada dia, de uma semana escolhida aleatoriamente, foi de: (Considere um $\alpha = 0{,}05$.)

Dia da semana	Número de acidentes
Segunda	20
Terça	10
Quarta	10
Quinta	15
Sexta	30
Sábado	20
Domingo	35

Chi Quadrado (χ^2) – Testes de Homogeneidade e de Independência

O que são?

Teste de Homogeneidade

Verificar se uma variável aleatória se comporta de modo similar, ou homogêneo, em várias subpopulações. Em outras palavras, em um teste de Chi Quadrado de homogeneidade podemos testar a afirmação de que diferentes populações têm a mesma proporção de indivíduos com alguma característica.

Teste de Independência

O teste de Chi Quadrado de independência é semelhante ao teste de Chi Quadrado de aderência, mas considera uma "lei oriunda da própria tabela de dados experimentais" a fim de avaliar se há ou não dependência entre duas variáveis. Quanto maior a dependência entre as duas variáveis, maior será o valor de χ^2. Quando as duas variáveis são independentes, o valor de χ^2 tende a zero.

Diferença entre os testes de <u>Homogeneidade</u> e de <u>Independência</u>: Hipóteses a serem testadas

<u>Teste de Homogeneidade</u>:

H0: o comportamento da variável é homogêneo nas subpopulações

Ha: o comportamento da variável não é homogêneo nas subpopulações

$$e_{i,j} = n_{linha\ i} \times \frac{total\ da\ coluna\ j}{total\ geral}$$

<u>Teste de independência:</u> distribuímos uma amostra de **N** elementos de "uma" população segundo as categorias da variável **A** e as categorias da variável **B**.

H0: A e B são variáveis independentes

Ha: As variáveis A e B não são independentes

$$e_{i,j} = \frac{n_{i.} \times n_{.j}}{n_{..}}$$

Exemplo

Em duas escolas diferentes, os estudantes são submetidos a um mesmo exame, em que A, B, C, D e E são as notas obtidas por eles:

	A	B	C	D	E
Escola 1	18	39	129	48	66
Escola 2	18	26	41	6	9

A distribuição das notas obtidas pelos estudantes é a mesma nas duas escolas? Considere um $\alpha = 0{,}05$.

Para estimar o valor do p, considere a tabela de Chi Quadrado fornecida em anexo.

<u>RESOLUÇÃO:</u>

1ª providência: Deixar claro quem são nossas hipóteses:

H0 = Hipótese nula = Hipótese de igualdade = A distribuição das notas é a mesma nas duas escolas

Ha = Hipótese alternativa = Hipótese da diferença = A distribuição das notas não é a mesma nas duas escolas

Quais são nossas variáveis?

<u>Variável Nota</u>, com cinco categorias: A, B, C, D e E.

<u>Variável Escola</u>, com duas subpopulações: Escola 1 e Escola 2.

2ª providência: verificar se todos os pré-requisitos para um *Chi*-Quadrado são respeitados:

Pelo menos 5 observações por casela → OK

Menos de 20% das caselas com ZERO → OK

3ª providência: encontrar os valores totais para linhas e colunas da tabela de frequência observada:

	A	B	C	D	E	***Total***
Escola 1	18	39	129	48	66	300
Escola 2	18	26	41	6	9	100
Total	36	65	170	54	75	400

4ª providência: calcular a frequência esperada, lembrando que para cada casela, a frequência esperada é igual a

$$Esperado = \frac{\sum valores\ coluna \times \sum valores\ linha}{N_{total}}$$

Segue a**Tabela valores esperados (segundo teste de homogeneidade):**

	A	B	C	D	E
Escola 1	300*(36/400 = 27	300*(65/400) = 48,75	300*(170/400) = 127,5	300*(54/400) = 40,5	300*(75/400) 56,25
Escola 2	100*(36)/400 = 9	100*(65/400) = 16,25	100*(170/400) = 42,5	100*(54/400) = 13,5	100*(75/400) 18,75

5ª providência: calcular o resíduo para cada casela, construindo a **Tabela de Resíduos (= Frequência observada – Frequência Esperada)**

	A	B	C	D	E
Escola 1	-9	-9,75	1,5	7,5	9,75
Escola 2	9	9,75	-1,5	-7,5	-9,75

6ª providência: cálculo do Chi Quadrado

$$\chi^2 = \frac{Resíduo_A^2}{Esperado_A} + \frac{Resíduo_B^2}{Esperado_B} + \frac{Resíduo_C^2}{Esperado_C} + \cdots + \frac{Resíduo_E^2}{Esperado_E}$$

$$\chi^2 = 3{,}00 + 1{,}95 + 0{,}02 + 1{,}39 + 1{,}69 + 9{,}00 + 5{,}85 + 0{,}05 + 4{,}17 + 5{,}07$$

$$\chi^2 = 32{,}19$$

Para o cálculo de graus de liberdade, temos que calcular:

$$g.l. = (N_{coluna} * N_{linha} - N_{coluna}) - (N_{coluna} - 1) = (N_{coluna} - 1) * (N_{linha} - 1)$$

Como trabalhamos com uma tabela 5x2, temos que o número de graus de liberdade é igual a:

$$g.l. = (N_{coluna} - 1)(N_{linha} - 1) = (5 - 1)(2 - 1) = 4 \times 1 = 4$$

Conforme a Tabela do Chi Quadrado, veremos que o valor de χ^2 com 4 grau de liberdade e p-valor = 0,05, esperaríamos um valor de 9,49 (χ^2 tabelado).

Como o valor calculado de $\chi^2_{(4)} = 32{,}19$ é BEM MAIOR do que 9.49, χ^2 tabelado crítico a p-valor = 0,05, podemos verificar também que mesmo se assumirmos p-valor < 0,001, onde $\chi^2_{(4)}$ tabelado = 18.47, o nosso $\chi^2_{(4)}$ calculado é ainda MAIOR. Como **χ^2 tabelado** é **MENOR** que **χ^2 calculado**, **<u>REJEITAMOS</u>** a hipótese nula mesmo a um $\alpha < 0{,}001$

7ª (e última) providência: Reportar por extenso

Para um nível de significância de 5%, rejeitamos a H0, ou seja, a distribuição das notas obtidas pelos estudantes não é a mesma nas duas escolas ($\chi^2_{(4)} = 32{,}19$; $p < 0{,}001$).

Exercício 1

Uma pesquisa classificou uma amostra de 800.000 alunos dos ensinos fundamental e médio (sendo 500.000 alunos do turno matutino e 300.000 alunos do turno vespertino) de acordo com o seguinte critério de rendimento escolar: ótimo, regular e péssimo, conforme a tabela a seguir:

Turno	Rendimento Ótimo	Rendimento Regular	Rendimento Péssimo
Matutino	80.000	300.000	120.000
Vespertino	150.000	60.000	90.000

O objetivo da pesquisa era avaliar se havia ou não dependência entre a variável "turno de estudos dos alunos dos ensinos fundamental e médio" e a variável "rendimento escolar". Verifique se houve associação entre as variáveis. Considere um $\alpha = 0{,}05$. Para estimar o valor do p, considere a tabela de Chi Quadrado fornecida em anexo.

Exercício 2

Realizou-se uma pesquisa com uma amostra de 400 frequentadores de um clube esportivo, sendo 160 mulheres e 240 homens, a fim de classificá-los de acordo com a modalidade esportiva preferida: vôlei, basquete ou tênis. Os dados coletados na pesquisa estão descritos na tabela a seguir:

Gênero	**Vôlei**	**Basquete**	**Tênis**
Mulheres	80	30	50
Homens	40	140	60

Verifique se há ou não dependência entre a variável “gênero” e a variável “modalidade esportiva preferida”. Considere um $\alpha = 0{,}05$. Para estimar o valor do p, considere a tabela de Chi Quadrado fornecida em anexo.

Exercício 3

Um pesquisador resolveu avaliar se a droga Prozac (fluoxetina) apresentaria efeitos benéficos no tratamento da anorexia, em pacientes que sofrem desse distúrbio. O Prozac é um inibidor seletivo da recaptação de serotonina, utilizado normalmente para combater sintomas de depressão, pânico, ansiedade, e sintomas obsessivos-compulsivos. Após o tratamento, observou-se se o quadro de anorexia seria superado (Sucesso), ou se haveria reincidência do distúrbio alimentar (Falha). Com base nos resultados abaixo, justifique se há associação entre o tipo de tratamento e a permanência do distúrbio alimentar. Considere um $\alpha = 0{,}05$.

	Anorexia	
Droga	Sucesso	Falha
Prozac	13	36
Placebo	14	30

Exercício 4

Um amante dos Três Patetas resolveu contar e dividir em categorias o número de tapas na cara sofrido por cada um dos Patetas ao longo de 199 episódios, originalmente criados para a televisão. Existe associação entre o número de tapas sofridos e quem o sofreu? Considere um $\alpha = 0{,}05$. Os resultados dessa análise estão apresentados na tabela a seguir:

	Pateta que recebeu o tapa		
Número de tapas	Curly	Shemp	Joe
0 a 10 tapas	49	34	10
11 a 20 tapas	36	21	5
21 a 30 tapas	7	14	5
Mais de 31 tapas	5	8	5

Exercício 5

Em um parque da cidade, um jovem estudante constatou um alto número de árvores doentes. Aparentemente, a doença se dá graças a um fungo que se instala na raiz das plantas e inicia um processo patológico que pode culminar com a morte da árvore. Com base nos registros de informação, justifique se existe uma associação entre a espécie de árvore e a presença da doença. Considere um $\alpha = 0{,}05$.

Árvore	***Fungo***	
	Infectada	**Saudável**
Ipê	110	10
Jacarandá	20	60

Teste Exato de Fisher

Em amostras pequenas o erro do valor de Qui-quadrado é alto e, portanto, o teste não é recomendável.

Ronald Fisher apresentou outro teste que permite calcular a ***probabilidade de associação das características*** que estão em análise, ou seja, a probabilidade de tais características serem independentes, quando o número total de dados é *pequeno*. Assim, em amostras pequenas deve-se executar esse teste, pois produz erro menor que o teste de Qui Quadrado.

De modo geral usa-se o Teste exato de Fisher quando: **(1)** o valor de *N < 20 ou* **(2)** *20 < N < 40* e a menor frequência esperada for menor que 5. As tabelas utilizadas são normalmente 2x2 (2 fatores discriminados e 2 fatores discriminantes):

		Fator Discriminado B		
		B1	B2	Totais de A
Fator Discriminante A	A1	O_{11}	O_{12}	N_{A1}
	A2	O_{21}	O_{22}	N_{A2}
	Totais de B	N_{B1}	N_{B2}	N_{TOTAL}

Nesse caso, a **Hipótese nula** é de que a frequência do Fator Discriminado **é independente** ao Fator Discriminante.

A nossa **hipótese alternativa** é de que a frequência do Fator Discriminado **depende** do Fator Discriminante

O teste exato de Fisher também pode ser **unicaudal** ou **bicaudal**. Para calcularmos um teste **unicaudal**, nós vamos retirando uma unidade de um lado dos fatores e acrescentando uma unidade ao outro fator, para determinar se a tendência da nossa hipótese alternativa (mais afastamento da nossa hipótese nula) é verdadeira (quanto mais A1 mais B1, por exemplo). Calculando o F0, depois o F1 e por fim o F2. O p-valor será igual a F0+F1+F2. Veja o procedimento **unicaudal** a seguir:

		Fator Discriminado B		
		B1	B2	Totais de A
Fator Discriminante A	A1	O_{11}	O_{12}	N_{A1}
	A2	O_{21}	O_{22}	N_{A2}
	Totais de B	N_{B1}	N_{B2}	N_{TOTAL}

Cálculo de F0:

$$F_0 = \frac{N_{A1}! \times N_{A2}! \times N_{B1}! \times N_{B1}!}{O_{11}! \times O_{12}! \times O_{21}! \times O_{22}! \times N_{Total}!}$$

Para cálculo de F1. Lembrando que H1 é de que **MAIOR** A1 → **MAIOR** B1 (logo, maior A2 maior B2):

		Fator Discriminado B		
		B1	B2	Totais de A
Fator Discriminante A	A1	O_{11}**+1**	O_{12} -1	N_{A1}
	A2	O_{21} -1	O_{22}**+1**	N_{A2}
	Totais de B	N_{B1}	N_{B2}	N_{TOTAL}

$$F_1 = \frac{N_{A1}! \times N_{A2}! \times N_{B1}! \times N_{B1}!}{(O_{11}+1)! \times (O_{12}-1)! \times (O_{21}+1)! \times (O_{22}-1)! \times N_{Total}!}$$

E F2:

		Fator Discriminado B		
		B1	B2	Totais de A
Fator Discriminante A	A1	O11+2	O12 -2	NA1
	A2	O21 -2	O22+2	NA2
	Totais de B	NB1	NB2	NTOTAL

$$F_2 = \frac{N_{A1}! \times N_{A2}! \times N_{B1}! \times N_{B1}!}{(O_{11}+2)! \times (O_{12}-2)! \times (O_{21}+2)! \times (O_{22}-2)! \times N_{Total}!}$$

Por fim calculamos o P-valor, que será igual a **$P = F_0 + F_1 + F_2$**

Se **P for menor que 0,05 (nível de significância: α)**, logo o nosso **teste é significativo** (rejeitamos a nossa hipótese nula), caso contrário, o nosso teste não é significativo. Se o nosso teste fosse bicaudal, a hipótese alternativa seria o afastamento mútuo dos valores, logo **$P_{bicaudal} = 2xP_{unicaudal}$.**

Exercício 1

De maneira geral, os doentes psiquiátricos podem ser classificados em psicóticos e neuróticos. Um psiquiatra realiza um estudo sobre os sintomas suicidas em duas amostras de 20 doentes de cada grupo. A nossa hipótese é que a proporção de psicóticos com sintomas suicidas é igual a proporção de neuróticos com estes sintomas (em um teste de independência, a hipótese nula seria, a presença ou ausência de sintomas suicidas é independente do tipo de doente envolvido). Assim, temos os dados resumidos na tabela a seguir.

		Tipo de Doente		Total
		Psicótico	Neurótico	
Sintomas Suicidas	Presente	2	6	8
	Ausente	18	14	32
Total		20	20	40